DEVELOPMENT OF COMPUTATIONAL RESOURCES FOR PREDICTING DISEASE RESISTANCE GENES AND MAPPING NGS-TRANSCRIPTS TO SECONDARY METABOLISM IN PLANTS

SYNOPSIS

Submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

In

Bioinformatics

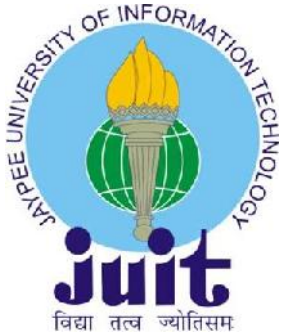
By

TARUN PAL

No. 116504

Under the supervision of

Prof. Rajinder Singh Chauhan



DEPARTMENT OF BIOTECHNOLOGY AND BIOINFORMATICS

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY

WAKNAGHAT-173234, HP, INDIA

FEBRUARY 2017

DEVELOPMENT OF COMPUTATIONAL RESOURCES FOR PREDICTING DISEASE RESISTANCE GENES AND MAPPING NGS-TRANSCRIPTS TO SECONDARY METABOLISM IN PLANTS

TARUN PAL

Introduction

Plants are in a constant onslaught from different pathogenic microorganisms including fungi, virus, bacteria, nematodes and oomycetes. To perceive such disease-related attacks plants possess disease resistance (R) genes, which have the ability to counter attack these pathogens. Plant disease resistance genes (R-genes) encode R-proteins which recognize proteins expressed by pathogens precise avirulence (Avr) genes [1,25].

Majority of R-genes are composed of nucleotide binding site (NBS) and a Leucine rich repeat (LRR) domain(s) referred to as (NBS-LRR) R-genes [1,2,3]. NBS-LRR are further classified as (TIR)-NBS-LRR (TNL) or non-TNL/coiled coil-NBS-LRR (CNL). In non-TIR-NBS-LRRs, the TIR domain may be substituted by a coiled-coil (CC) domain, defining CC-NBS-LRR proteins.

Considering the importance of disease resistance genes, few computational resources for R genes have been developed in the past. The most popular resource PRGdb (Plant Resistance Gene database) is a dedicated repository which hosts 112 reference R proteins [4]. Early prediction of R proteins can accelerate the process of genetic improvement and breeding programs for the development of disease-resistant varieties [5]. The existing methods are time-consuming and require the capability of extensive data processing. To overcome this problem, there is need for alignment and domain-independent methods. Since the above methods are based on high sequence similarity, certain low similarity proteins cannot be predicted through these methods.

Furthermore, we have also performed computational analysis of transcriptomes of two important medicinal herbs *Aconitum heterophyllum* and *Swertia chirayita* to unravel genes responsible for secondary metabolism. Next-generation sequencing (NGS) has provided unprecedented opportunities for high throughput research on medicinal plants, whose genome/transcriptome datasets were still not available. The medicinal herbs, *Aconitum heterophyllum* and *Swertia chirayita* have been extensively used in various herbal drug formulations resulting in their excessive utilization; thereby, placing them in high value endangered category. *Aconitum heterophyllum* Wall (Ranunculaceae), popularly known as Atis, is a biennial herb distributed at altitudes of 2400–3600 m, in sub-alpine and alpine regions of the Himalyas. The tuberous roots of *A. heterophyllum* contain nontoxic alkaloids like atisine, aconitine, hetidine and heterophyllinine [6,7]. Whereas, *Swertia chirayita* (*S. chirayita*) commonly known as ‘Chirata’, belonging to family Gentianaceae, is endemic to temperate Himalayas found at an altitude of 1200-3000 m [8]. It processes diverse pharmacological properties due to the presence of large number of secondary metabolites, including swertiamarin, mangiferin and amarogentin.

Transcriptome analysis can reveal important molecular components associated with the biosynthesis and accumulation of secondary metabolites in medicinal herbs within a stipulated time frame.

However, there is no information existing as of today on the transcriptome resources of *A. heterophyllum* and *S. Chirayta* endangered plant species. Due to lack of transcriptomic information for these species there exists a gap not only for discovering candidate genes involved in secondary metabolites production, but also for understanding molecular basis of various biological processes. Therfore viewing these research gaps, the transcriptomes of these two important species were assembled, annotated and characterized for the first time to decipher molecular components contributing to biosynthesis and accumulation of secondary metabolites.

Keeping in view lack of a machine learning tools for prediction of disease resistance proteins in plant species and immense importance of medicinal plants, gaps in knowledge on biosynthesis of secondary metabolites the present research was carried with the following objectives:

Objective 1: Development of a machine learning based computational tool for prediction of disease resistance proteins in plants

Objective 2: Computational analysis of *Aconitum heterophyllum* and *Swertia chirayita* transcriptomes to unravel genes responsible for secondary metabolism

Objective 1: ***Development of a machine learning based computational tool for prediction of disease resistance proteins in plants***

Methods

A collection of 112 reference *R* proteins from 25 plant species was downloaded from the PRGdb database [10] to serve as a positive dataset belonging to seven different domain classes (Table 1). A total of 119 non-*R* protein sequences were selected to act as a negative dataset. The flow diagram for the methodology adopted in the current tool is given in Fig. 1.

Table 1 Reference proteins and their domain classes

Sr no.	Reference Domain classes	Number of reference Proteins
1	TNL (TIR-NBS-LRR)	16
2	CNL (CC-NBS-LRR)	50
3	NL (NBS-LRR)	4
4	RLP (Receptor-like proteins)	12
5	RLX (Receptor-Like Kinases)	10
6	TN (TIR-NBS)	1
7	Others (conferring resistance)	19
	Total	**112**

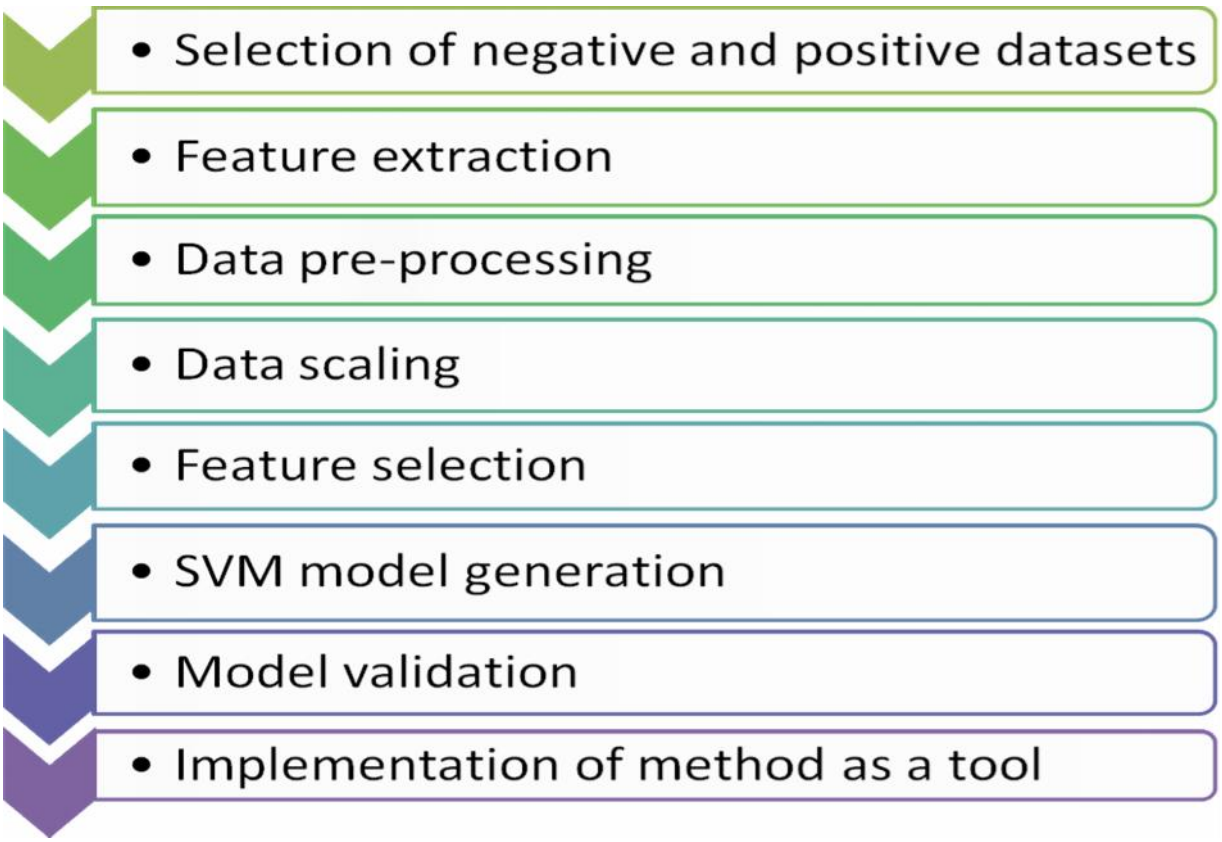

• Selection of negative and positive datasets
• Feature extraction
• Data pre-processing
• Data scaling
• Feature selection
• SVM model generation
• Model validation
• Implementation of method as a tool

Fig. 1 Flow diagram of the methodology used in DRPPP

Division of dataset into training and test sets

Selected positive and negative datasets were further divided into training and test sets. Of these, 80% of random sequences were taken as training set and the remaining 20% as test set (Table 2).

Table 2 Data set distribution into training, test, positive and negative classes

Sr no.	Data set	Positive	Negative	Total
1	Training	90	96	186
2	Test	22	23	45
	Total	112	119	

Feature Extraction

The Protr package in R language is capable of generating descriptors (features) explicitly for bioinformatics and chemogenomics research on protein sequences [9]. Important descriptors include:

1. Amino Acid Composition: It describes fraction of individual amino acid type within total number of amino acids in protein. AAC descriptors are defined as:

$$f(r) = \frac{N_r}{N} \quad r = 1, 2, ..., 20.$$

Where N_r is the number of the amino acid and N is the length of the sequence.

2. Autocorrelation: It describes amino acid properties along the sequence. It is defined as:

$$AC(d) = \sum_{i=1}^{N-d} P_i P_{i+d} \quad d = 1, 2, \ldots, \text{nlag}$$

Where d is called the lag of the autocorrelation.

3. Composition / Transition / Distribution: The composition is referred as the global percent for each encoded class in the sequence. Transition from one class to another is the percent frequency with which one class is followed by other or second class is followed by first in encoded sequence. Distribution descriptor states the distribution of each attribute in the sequence.

$$C_r = \frac{n_r}{n} \quad r = 1, 2, 3$$

Where n_r is the number of amino acid type r in the encoded sequence and N is the length of the sequence.

Transition is defined as:

$$T_{rs} = \frac{n_{rs} + n_{sr}}{N - 1} \quad rs = '12', '13', '23'$$

where n_{rs} , n_{sr} is the numbers of dipeptide encoded as "*rs*" and "*sr*" respectively in the sequence and *N* is the length of the sequence.

4. Conjoint Triad Descriptors: It considers the properties of vicinal amino acids and properties of one amino acid, thereby considering three continuous amino acids as a unit. Generates 20 plus discrete numbers to represent a protein. It is described as:

$$d_i = \frac{f_i - \min\{f_1, f_2, ..., f_{343}\}}{\max\{f_1, f_2, ..., f_{343}\}}$$

Where the numerical value of d_i for each protein ranges from 0 to 1

5. Pseudo-Amino Acid Composition (PAAC): They generate 20 + λ discrete numbers to represent a protein. PAAC descriptors are calculated using these functions extractAPAAC() and extractPAAC() and defined as:

$$H_1(i) = \frac{H_1^0(i) - \frac{1}{20}\sum_{i=1}^{20} H_1^0(i)}{\sqrt{\frac{\sum_{i=1}^{20}[H_1^0(i) - \frac{1}{20}\sum_{i=1}^{20} H_1^0(i)]^2}{20}}}$$

6. Quasi-sequence-order Descriptors: Physicochemical distance between amino acids, a set of sequence-order coupling numbers was introduced to reflect the sequence order effect, or in a rigorous term, the quasi sequence-order effect. They are derived from the distance matrix between the 20 amino acids. This descriptor and is calculated by:

$$X_r = \frac{f_r}{\sum_{r=1}^{20} f_r + \sum_{d=1}^{\text{maxlag}} {}^2{}_d} \quad r = 1, 2, 3, ..., 20$$

Pre-processing

Data scaling was introduced for pre-processing of data to produce good accuracy, where scaling of all the features was achieved between 0 and 1.

Feature selection

It is used for reducing the computation time and computational complexity. Feature selection was performed using Fselect script from the libSVM package [10].

Model evaluations

Evaluation of test datasets

The performance of the model was evaluated using test datasets. The data files from the test datasets were normalized according to the SVM model requirements i.e. between 0 and 1.

10-fold cross-validation on the training dataset

In this approach, training data was divided into 10 equal subsets, followed by training of nine datasets, test on one dataset, repetition for ten times, and then mean accuracy was calculated [11].

Performance of the support vector machine model using statistics

The following equations were used for calculating sensitivity, specificity, accuracy and Matthew correlation coefficient (MCC):

1. Sensitivity: It is defined as the percentage of R proteins that are correctly predicted.

Sensitivity = TP/TP+FN*100

2. Specificity: It is the percentage of non R proteins that are correctly predicted as non-R proteins.

Specificity = TN/TN+FP*100

3. Accuracy: It is the percentage of correct predictions from total number of predictions.

Accuracy =TP+TN/TP+FP+TN+FN*100

4. Matthews correlation coefficient (MCC): It is a combined measure of both sensitivity and specificity.

MCC= (TP*TN)-(FN*FP)/ (sqrt(TP+FN)*(TN+FP)*(TP+FP)*(TN+FN)

ROC curve

The ROC plot and AUC were calculated for the SVM classifier used in the study using libSVM package by ranking the decision values.

Tool and website implementation

The model predicted in this study is implemented as a free standalone tool 'DRPPP' to predict *R* genes and is freely accessible on http://14.139.240.55/NGS/download.php.

Results and Discussion

Development of disease resistance has always been one of the major objectives for any crop improvement program and early identification of *R* genes can be a major step towards achieving this goal. Databases and tools such as PRGdb and DRAGO are available in the literature, but they rely on similarity or domain prediction methods only, neglecting the relevance of low similarity proteins and also hindering the identification of new or novel *R* genes, which cannot be predicted through these tools.

10-fold cross-validation was performed on the training dataset using the radial basis function (RBF) kernel. The numbers of predicted sequences from TP, TN, FP and FN were 18, 23, 0 and 4, respectively from a total of 45 test datasets. The performance of the module is provided in Table 3.

Table 3 Performance of support vector machine model on dataset

Sr. no.	Model	Sensitivity (%)	Specificity (%)	Accuracy (%)	MCC
1	SVM	81.82	100	91.11	0.83

Ten-fold cross validation was performed to optimize SVM parameter using the radial basis function (RBF) kernel. Training dataset cross validation accuracy was achieved to be 85.4839%. The Receiver Operating Characteristic (ROC) curve was used to show the trade-off between true positive rate (sensitivity) and false positive rate (specificity) for all possible values (Fig. 2).

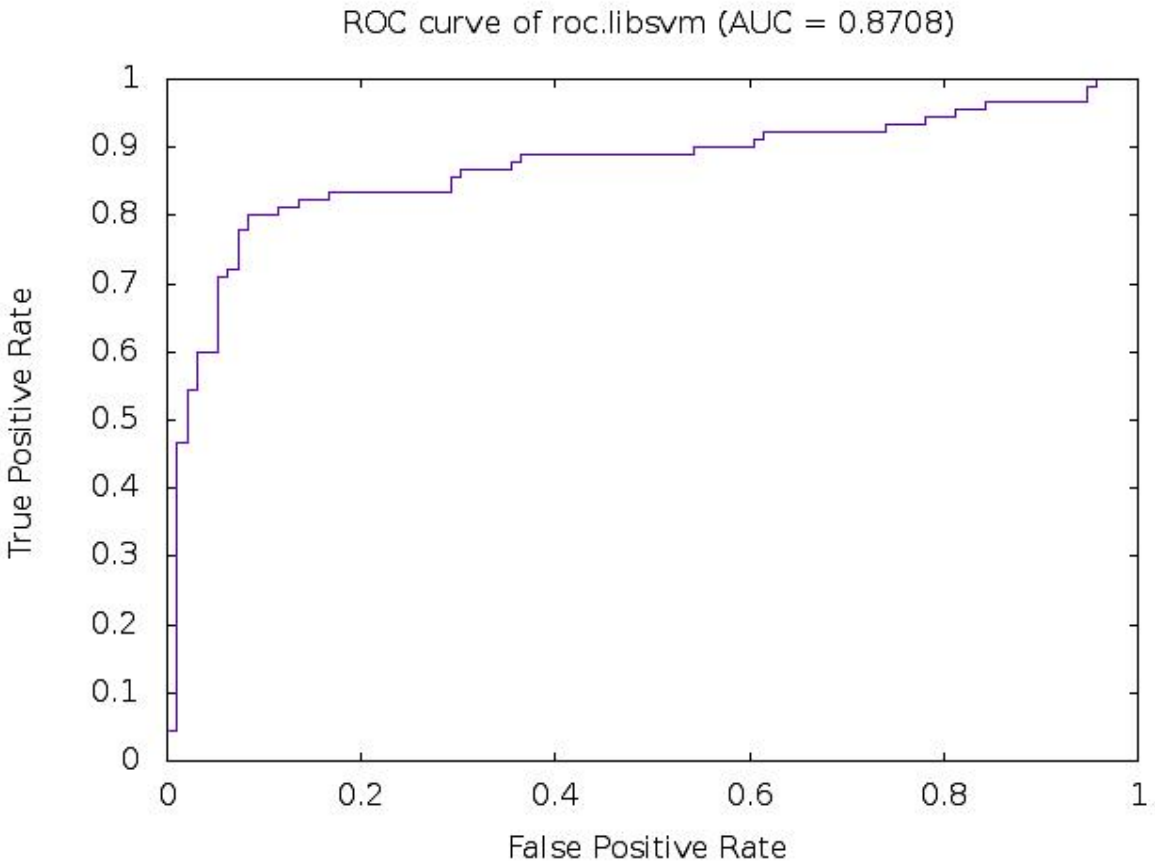


Fig. 2 ROC Curve for Binary SVM classifier: ROC plot depicts relative trade-offs between true positive and false positives.

The test dataset of DRPPP was run on NBSPred, DRPPP detected 91.11 % of the sequences as *R* protein, whereas NBSPred detected 88.88% indicating more exploration potential of DRPPP for predicting *R* proteins.

The 'DRPPP' is a machine learning computational tool, which can identify important *R* proteins in plants. Moreover, it builds the model based on seven descriptor groups with sixteen descriptors, including compositional frequencies descriptors used by NBSPred. All these parameters indicated the high reliability of the current tool. It is intended that DRPPP will provide a new research perspective for identification of *R* genes by observing a large set of protein datasets to produce an unbiased smaller set of more precisely predicted *R* genes candidates.

Conclusion

The influential features for predicting *R* genes are not known completely, therefore, a set of all important features was generated through feature extraction techniques. DRPPP can efficiently predict *R* protein with highest accuracy (91.11%) as compared to other existing tools.

Objective 2: ***Computational analysis of Aconitum heterophyllum and Swertia chirayita transcriptomes to unravel genes responsible for secondary metabolism***

Material and Methods

Medicinal plant species

We have selected two medicinal herbs (*A. heterophyllum* and *S. chirayita*) from North-Western Himalayas having high therapeutic values. Different tissues were selected on the basis of differential amounts of secondary metabolite biosynthesis and accumulation for transcriptome generation and their subsequent analysis.

Transcriptomes generation

A total of four transcriptomes two from each plant species (*A. heterophyllum* and *S. chirayita*) were generated and analyzed to decipher the molecular components associated with the secondary metabolite synthesis. For transcriptomes generation, root (AHSR) and shoot tissues (AHSS) of *A. heterophyllum*; and tissue cultured (SCTC) and greenhouse grown (SCFG) plants of *S. chirayita* were used in the current research work (Fig. 3).

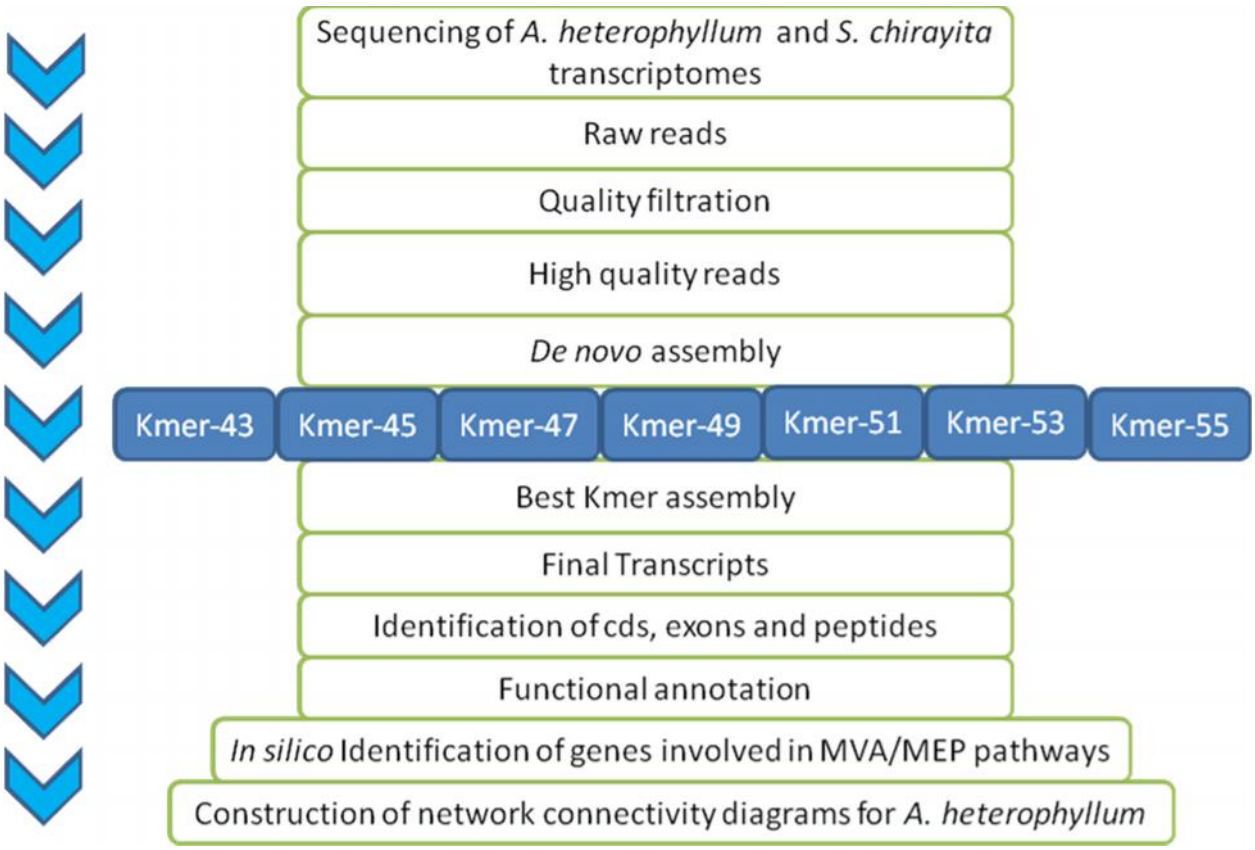


Fig. 3 Work flow of *de novo* whole transcriptome analysis for root (AHSR) and shoot (AHSS) transcriptomes of *A. heterophyllum* and tissue culture (SCTC) and greenhouse grown (SCFG) plant transcriptomes of *S. chirayita*.

De novo assembly

Once the sequencing has been done, useful and meaningful information needs to be extracted from the raw data. GENSCAN gene prediction webserver [21] was used for identification of coding sequence (CDS), exons and peptides.

Functional annotation, GO mapping and COG analysis

To further annotate assembled transcripts, the BLAST2GO program [12] was used to retrieve GO annotation, which differentiates according to molecular function, biological process and cellular component ontologies (http://www.geneontology.org/). In-house developed scripts were used to align the assembled sequences for all the four transcriptomes to the COG database [17] so as to predict and classify possible functions.

```
open(F1,"chech") or die "file not found";
open(F2,"whog.txt") or die "file not found";
while(<F1>)
{
chomp($_);
@a=split(/>/,$_);
push(@b,$a[1]);
}
while(<F2>)
{
	if($_=~ /COG/)
	{

		@d=split(/\s/,$_);
		$count=1;

	}
		if ($count==1)
			{
				foreach$x(@b)
				{
					if (/$x/)
					{

						print "$x\t$d[0]\n";
					}
				}
			}
}
```

Fragment mapping and transcript abundance prediction

Transcript quantification for the *de novo* assembled sequences was carried out using RSEM approach [18] which assesses the transcript abundances based on the mapping of RNA-Seq reads to the assembled transcriptome.

Pathway mapping using KEGG in A. heterophyllum and S. chirayita transcriptomes

The KEGG gene database has an edge over other databases, as it acts as a single resource for cross-species representation by assigning KEGG orthology (KO) to all available genomes [19]. These ortholog candidates were grouped by KO, and each KO group was assigned score. Finally, KO was ranked in accordance with the highest calculated a scores.

Functional classification and network connectivity diagrams

All gene IDs mentioned in the annotation file were mined for all the transcripts sequences from both the plant species. Retrieved gene IDs were mapped onto their respective biological processes in NCBI BioSystems database [20]. Network connectivity diagrams [24] were drawn for isoquinoline alkaloids biosynthesis in *A. heterophyllum* root transcriptome using perl-tk.

```
use GD;
use GD::Arrow;
open(F4,"pathway_information.txt");
while(<F4>)
	{
		chomp($_);
		@w=split(/\t/,$_);
		$s{$w[0]}=$w[2];
	}
close(F4);
opendir F1,"./information/";
@wa=readdir F1;
closedir(F1);
mkdir "./information1_long";
foreach $x8(@wa)
		{
			if($x8 ne '.'  && $x8 ne '..' )
				{
					$count=0;
					undef %count1,%count2,%count7;
					open(F2,"./information/".$x8);
					open(F5,">./information1_long/".$x8."\.jpeg");
					while(<F2>)
						{
							$count7{$count}=0;
							chomp($_);
							@w1=split(/\t/,$_);
							foreach $x(@w1)
								{
									@w=split(/[|]/,$x);
```

```
if($count1{$count}{$w[1]} ne 1 ){

$count1{$count}{$w[1]}=1;

$name{$count}{$count7{$count}}=$w[1];

$count7{$count}=$count7{$count}+1;

$count3{$count}{$w[1]}=0;

$count2{$count}{$w[1]}{$count3{$count}{$w[1]}}=$w[1];

$count5{$count}{$w[1]}{$count3{$count}{$w[1]}}=$w[0];

$count3{$count}{$w[1]}=$count3{$count}{$w[1]}+1;
                                                                    }
                                                                    else
                                                                    {

$count2{$count}{$w[1]}{$count3{$count}{$w[1]}}=$w[1];

$count5{$count}{$w[1]}{$count3{$count}{$w[1]}}=$w[0];

$count3{$count}{$w[1]}=$count3{$count}{$w[1]}+1;}
                                                                }
                                                        $count=$count+1;
                                                }

                                        open(F2,">positions.txt");
                                        for($i=0;$i<$count;$i++)
                                                {

                                                for($i1=0;$i1<$count7{$i};$i1++)
                                                {

$position1{$i}{$i1}{x}=50+($i*540);
$position1{$i}{$i1}{y}=50+($i1*80);
print F2 $i." ".$i1." ".$position1{$i}{$i1}{x}." ".$position1{$i}{$i1}{y}."\n";

$position2{$i}{$i1}{x}=50+($i*540);

$position2{$i}{$i1}{y}=$position1{$i}{$i1}{y}+25;
$position3{$i}{$i1}{x}=$position1{$i}{$i1}{x}+400;

$position3{$i}{$i1}{y}=$position1{$i}{$i1}{y}+25;
print F2 $i." ".$i1." ".$position1{$i}{$i1}{x}." ".$position1{$i}{$i1}{y}."
".$position2{$i}{$i1}{x}." ".$position2{$i}{$i1}{y}." ".$position3{$i}{$i1}{x}."
".$position3{$i}{$i1}{y}."\n";
                                                                                }
                                                        }

                                        close(F2);
# create a new image
$im = new GD::Image(6500,3000);
```

```
   # allocate some colors
   $white = $im->colorAllocate(255,255,255);
   $black = $im->colorAllocate(0,0,0);
   $red = $im->colorAllocate(255,0,0);
   $blue = $im->colorAllocate(0,0,255);

# make the background transparent and interlaced
   $im->transparent($white);
   $im->interlaced('true');

        for($i1=0;$i1<$count;$i1++)
                {
                        for($i2=0;$i2<$count7{$i1};$i2++)
                                {
                                        $x1=$position1{$i1}{$i2}{x};
                                        $y1=$position1{$i1}{$i2}{y};
                                        $x2=$x1+400;
                                        $y2=$y1+50;

                                    if($x1 eq 50 && $y1 eq 50 )
                                        {
                                                $im>rectangle($x1,$y1,$x2,$y2,$red);
                                        }
                                    else
                                        {
                                                $im>rectangle($x1,$y1,$x2,$y2,$black);
                                        }

        $im>string(gdMediumBoldFont,($x1+5),($y1+5),$s{$name{$i1}{$i2}},$red);
                                }
                                }
        my $width = 1;
        my @arrow;
        $k=0;
        for($i=1;$i<$count;$i++)
                {
                for($i1=0;$i1<$count7{$i};$i1++)
                        {
                                for($i2=0;$i2<$count3{$i}{$name{$i}{$i1}};$i2++)
                                        {

                                        for($i3=0;$i3<$count7{($i-1)};$i3++)
                                                {
                                                        for($i4=0;$i4<$count3{($i-1)}{$name{($i-
1)}{$i3}};$i4++)
                                                                                {
if($count5{$i}{$name{$i}{$i1}}{$i2} eq $count2{$i-1}{$name{$i-1}{$i3}}{$i4} )
                                                                                                { #print
$count5{$i}{$name{$i}{$i1}}{$i2}." ".$count2{$i-1}{$name{$i-1}{$i3}}{$i4}."\n";
        $x2=$position3{($i-1)}{$i3}{x};
        $y2=$position3{($i-1)}{$i3}{y};
        $x1=$position2{$i}{$i1}{x};
        $y1=$position2{$i}{$i1}{y};
```

```
$arrow[$k] = GD::Arrow::Full->new(
                        -X1    => $x1,
                        -Y1    => $y1,
                        -X2    => $x2,
                        -Y2    => $y2,
                        -WIDTH => $width,
                    );
$im->filledPolygon($arrow[$k],$blue);
$k=$k+1;                                                    }

                                                            }
                                                }
                                            }
                    }
        }

binmode F5;
print F5 $im->jpeg;
close(F5);
undef
%count7,%count1,%count2,%count3,%count5,%name,%position,%position1,%position2,%position3,
$im;
        }
}
```

Website development and platform used

The website for free accessibility of detailed results corresponding to the analysis done was developed using HTML and PHP in combination with custom Java programs and Perl scripts (Fig. 4).

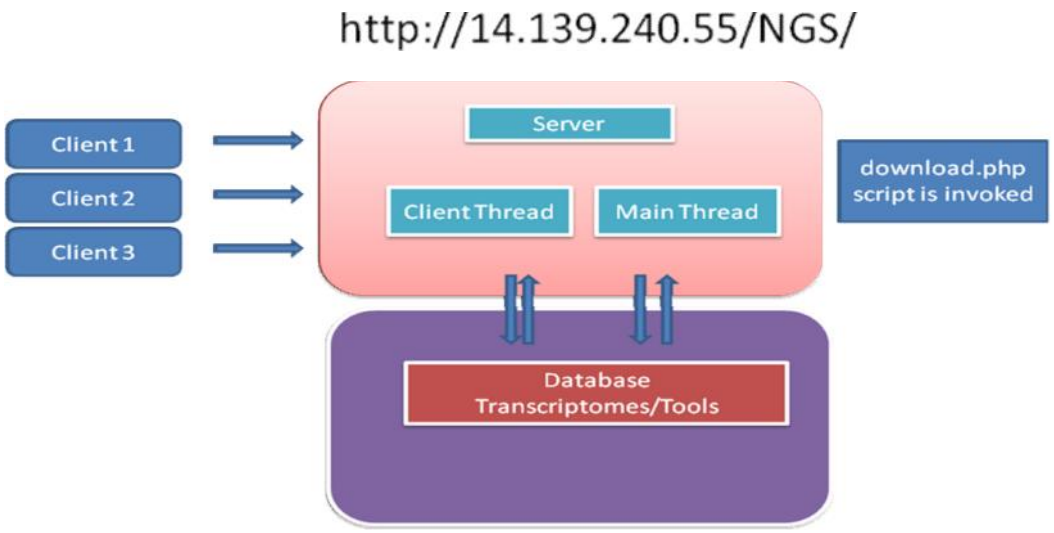


Fig. 4 NGSAR- (Next Generation Sequencing and Analysis Resources) server's architecture

Results and discussion

The transcriptome data provide a valuable resource for rapid elucidation of pathways, identification and characterization of biosynthesis pathway genes, pathway mapping and construction of network connectivity diagrams [13,16].

Illumina paired-end sequencing and de novo sequence assembly

Based on the assembler statistics, AHSR generated 75,548 transcripts, N50 of 1059 bp. Similarly, AHSS generated 39,100 transcripts, N50 of 1059 bp (Table 4). A total of 57,460 assembled transcripts were obtained for greenhouse *S. chirayita* sample (SCFG) (Table 5). Similarly, for tissue cultured *S. chirayita* sample (SCTC), a total of 43,702 transcripts were obtained. The difference in the number of transcripts generated for root and shoot *A. heterophyllum* samples may be due to differential accumulation condition for secondary metabolites and may be due to their differential growth conditions in *S. Chirayita.*

Table 4 Transcriptome assembly statistics for roots (AHSR) and shoots (AHSS) samples of *A. heterophyllum*

Description	Root Transcriptome	Shoot Transcriptome
Best kmer	kmer-51	kmer-43
Number of transcripts	75,548	39,100
Transcript N50	1,059	1,239

Table 5 Assembly statistics for transcriptomes from greenhouse (SCFG) and tissue culture (SCTC) samples of *S. chirayita*

Description	SCFG Transcriptome	SCTC Transcriptome
Best Kmer	Kmer-51	Kmer-47
Number of transcripts	57,460	43,702
Transcript N50	1,700	1,629

Functional annotation of transcripts

The transcripts were searched against NCBI non-redundant (nr) using BLASTX program with an E-value threshold of 1e-5. A total of 46,850 sequences had significant BLAST hits with nr protein database in AHSR, whereas for AHSS, 36,326 sequences showed significant BLAST hit with nr database. We applied GenScan and predicted 34,424 CDS, 41,700 exons and 34,424 peptides for AHSR and 23,149 CDS, 27,906 exons and 23,149 peptides for AHSS (Table 6).

For *Swertia* transcriptomes, Blastx resulted in the annotation of 50,795 transcripts with significant BLAST hits for SCFG. For SCTC 39,150 sequences showed significant BLAST

hit with nr database. GenScan prediction tool yielded 35,493 CDS, 44,277 exons and 34,493 peptides for SCFG and 26,349 CDS, 31,757 exons and 31,757 peptides for SCTC (Table 7).

A large number of transcripts could be matched with unique known proteins in public databases, which implied that the Illumina sequencing yielded a substantial fraction of unique genes from *A. heterophyllum* and *S. chirayita* transcriptomes.

Table 6 Prediction summary of CDS, exons and peptides in root and shoot transcriptomes of *A. heterophyllum*

Description	Root transciptome	Shoot transciptome
Number of CDS	34,424	23,149
Number of exons	41,700	27,906
Number of peptides	34,424	23,149

Table 7 Prediction summary of cds, exons and peptides in greenhouse (SCFG) and tissue culture (SCTC) transcriptomes of *S. chirayita*

Description	SCFG transcriptome	SCTC transcriptome
Number of cds	35, 493	26, 349
Number of exons	44, 277	31, 757
Number of peptides	35, 493	31, 757

Functional classification by GO and COG

The BLAST2GO program was used to obtain GO annotation for the transcripts annotated by the non-redundant database. Transcripts were assigned to GO classes with 27,596 and 12,340 functional terms for AHSR and AHSS, respectively. As shown in Fig. 5, the assignments to the molecular function made up the majority (11,996, 43.47 % for AHSR; 5665, 45.90 % for AHSS).

Similarly, for *Swertia* transcriptomes 18,090 and 2,102 functional terms were yielded for SCFG and SCTC, respectively. These transcripts were further classified into three major categories, the assignments to the molecular function made up the majority (8,278 and 45.76% for SCFG; 1,106 and 52.61% for SCTC) (Fig. 6). Among three GO categories, cell, catalytic and metabolic process were the most abundant classes in cellular component, molecular function and biological processes, respectively. These observations were in agreement with previously assigned GO terms in *Arabidopsis thaliana, Picrorhiza kurrooa* and *Medicago truncatula* [13, 14].

All transcripts were aligned to the COG database to predict and classify plausible functions. A total of 16,604 and 9,398 transcripts of AHSR and AHSS, respectively, were assigned to COG classifications. Among the 24 COG IDs, general function prediction was the largest

group in both tissue samples (2,870, 17.28 % in AHSR and 1,786, 19.00 % in AHSS) (Fig. 7). More importantly, 563 (3.39 %) and 385 (4.09 %) transcripts were classified in the group of secondary metabolites biosynthesis in *A. heterophyllum* from AHSR and AHSS, respectively. In the similar manner, we categorized Swertia transcripts into 24 functional classes, where the most frequent functional category was observed to be "general function prediction (symbol R)" for both tissue samples (2,405 in SCFG and 1,820 in SCTC) (Fig. 8), etc., which was similar to the study done on Taxodium [15]. Furthermore, 528 and 313 transcripts of SCFG and SCTC, respectively, were classified into "secondary metabolites biosynthesis, transport and catabolism (category symbol Q)" in *S. chirayita*.

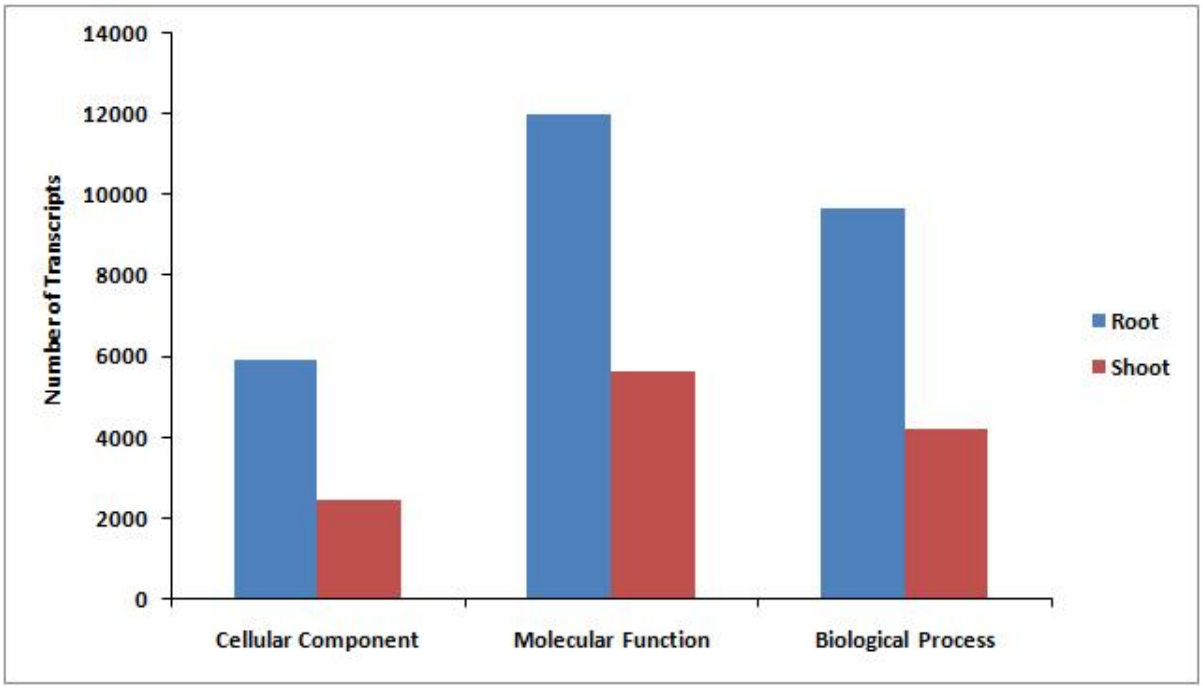


Fig. 5 GO distribution for *A. heterophyllum* root and shoot transcriptomes

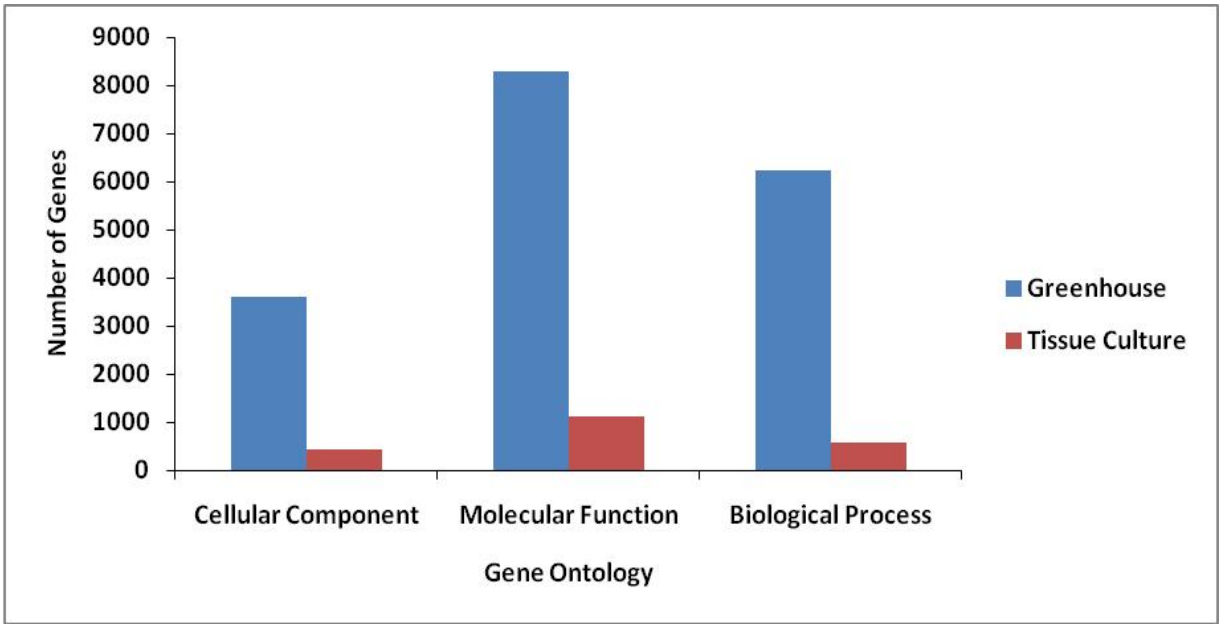


Fig. 6 Distribution of GO annotated transcripts in greenhouse and tissue culture transcriptomes of *S. chirayita*

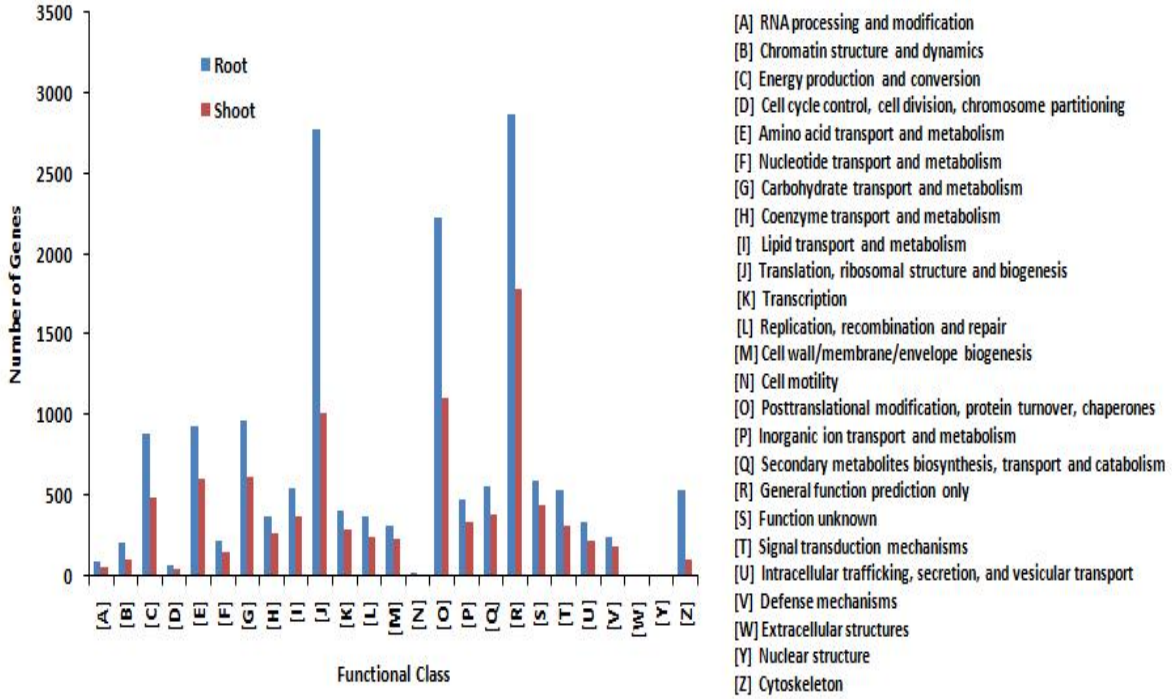


Fig. 7 COG classification of *A. heterophyllum* root and shoot transcriptomes

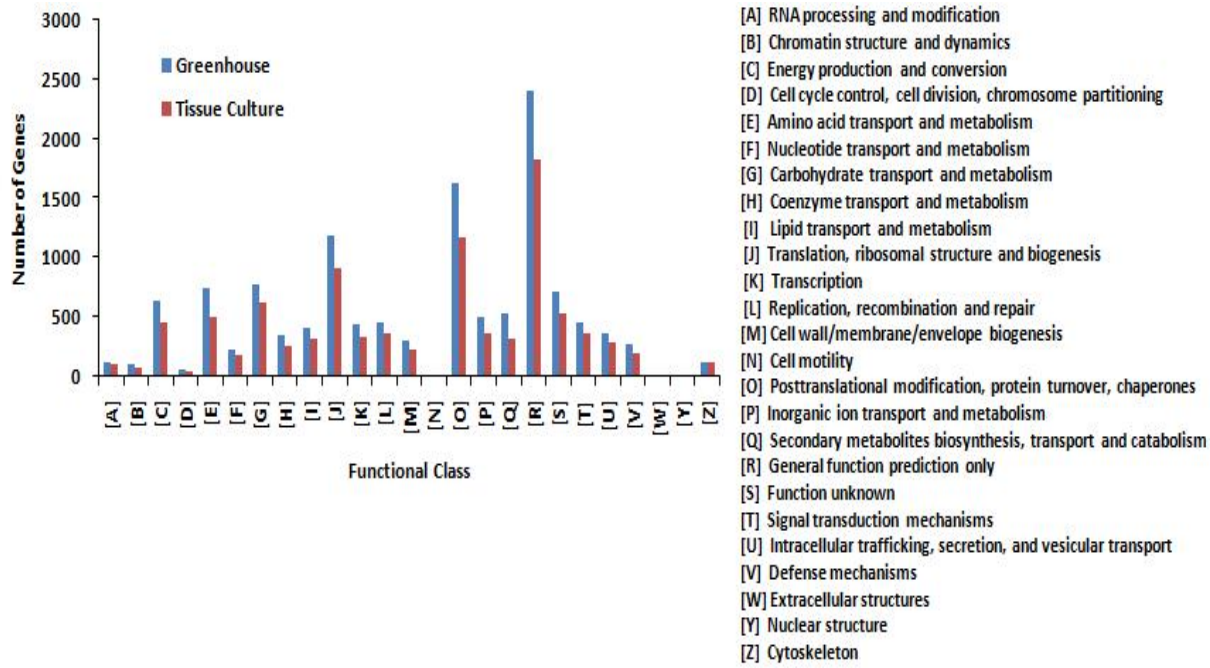

3000
2500
2000
1500
1000
500
0
Number of Genes
Greenhouse
Tissue Culture
[A] [B] [C] [D] [E] [F] [G] [H] [I] [J] [K] [L] [M] [N] [O] [P] [Q] [R] [S] [T] [U] [V] [W] [Y] [Z]
Functional Class
[A] RNA processing and modification
[B] Chromatin structure and dynamics
[C] Energy production and conversion
[D] Cell cycle control, cell division, chromosome partitioning
[E] Amino acid transport and metabolism
[F] Nucleotide transport and metabolism
[G] Carbohydrate transport and metabolism
[H] Coenzyme transport and metabolism
[I] Lipid transport and metabolism
[J] Translation, ribosomal structure and biogenesis
[K] Transcription
[L] Replication, recombination and repair
[M] Cell wall/membrane/envelope biogenesis
[N] Cell motility
[O] Posttranslational modification, protein turnover, chaperones
[P] Inorganic ion transport and metabolism
[Q] Secondary metabolites biosynthesis, transport and catabolism
[R] General function prediction only
[S] Function unknown
[T] Signal transduction mechanisms
[U] Intracellular trafficking, secretion, and vesicular transport
[V] Defense mechanisms
[W] Extracellular structures
[Y] Nuclear structure
[Z] Cytoskeleton

Fig. 8 Distribution of COG classified transcripts of greenhouse and tissuecultue transcriptomes of *S. chirayita*

Mining transcriptomes for genes involved in MVA/MEP biosynthesis pathways

All the key genes of mevalonate (MVA) and non-mevalonate (MEP) pathways involved in the biosynthesis of aconites in *A. heterophyllum* [22] were computationally mined using in-house developed perl script. We identified 15 enzymes from in-house generated transcriptomes namely *ACTH*, *HMGS*, *HMGR*, *MVK*, *PMK*, *MVDD*, *IPP*, *GDPS*, *DXPS*, *DXPR*, *ISPD*, *ISPE*, *MECPS*, *HDS*, *ISPH* for MVA and MEP pathways in AHSR and AHSS from nr database using BLASTX [23]. The analysis of gene expression for two tissues of *A. heterophyllum* was estimated using FPKM approach (Tables 8, Fig. 6).

Similarly, in case of Swertia, SCFG and SCTC transcriptomes were mined for genes involved in the secondary metabolites biosynthesis pathways. The transcript abundance analysis for MVA/MEP revealed that most of the genes (9 genes) showed higher transcript abundance in SCFG compared to SCTC transcriptomes, thereby suggesting their significant contribution in secondary metabolites production in *S. chirayita* (Table 9, Fig. 10).

Table 8 Transcript abundance of MVA and MEP pathways genes identified in *A. heterophyllum* root and shoot transcriptomes

Sr. No	Gene Name	E EC. No.	AHSR Transcript ID(s)	AHSS Transcript ID(s)	FPKM Root	FPKM Shoot
1.	*DXPS*	2.2.1.7	Transcript_5409	Transcript_1017	57.68	176.67
2.	*DXPR*	1.1.1.267	Transcript_1136	Transcript_7612	48.54	102.02
3.	*ISPD*	2.7.7.60	Transcript_10174	Transcript_27344	23.7	96.94
4.	*ISPE*	2.7.1.148	Transcript_64620	Transcript_11462	19.13	38.73
5.	*MECPS*	4.6.1.12	Transcript_1824	Transcript_3605	115.13	403.23
6.	*HDS*	1.17.7.1	Transcript_8398	Transcript_7021	**105.58**	90.12
7.	*ISPH*	1.17.1.2	Transcript_30192	Transcript_4360	20.09	296.29
8.	*ACTH*	2.3.1.9	Transcript_51038	Transcript_11170	21.43	22.22
9.	*HMGS*	2.3.3.10	Transcript_5690	Transcript_5417	16.25	84.79
10.	*HMGR*	1.1.1.34	Transcript_3795	Transcript_2381	**143.06**	48.26
11.	*MVK*	2.7.1.36	Transcript_19736	Transcript_14392	**14.12**	1.92
12.	*PMK*	2.7.4.2	Transcript_6636	Transcript_4131	4.31	26.54
13.	*MVDD*	4.1.1.33	Transcript_63336	Transcript_33574	**14.31**	12.28
14.	*IPP*	5.3.3.2	Transcript_10661	Transcript_1345	112.15	154.8
15.	*GDPS*	2.5.1.84	Transcript_52592	Transcript_5178	8.47	66.33

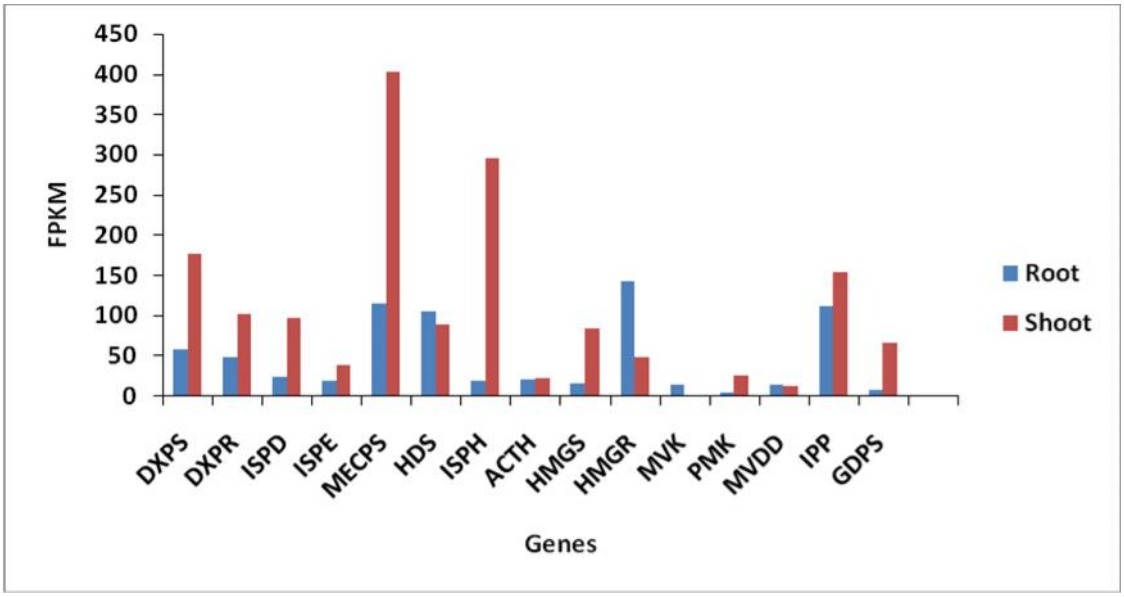


Fig. 9 Transcript abundance of MVA and MEP pathways genes in root versus shoot transcriptomes

Table 9 Transcript abundance of MVA and MEP pathways genes identified in greenhouse and tissue culture samples of *S. chirayita*

Sr. No	Gene Name	E.C. No.	SCFG Transcripts ID(s)	SCTC Transcript ID(s)	FPKM SCFG	FPKM SCTC
1.	*AACT*	2.3.1.9	transcript_54602	transcript_32680	76.94	82.11
2.	*HMGS*	2.3.3.10	transcript_520	transcript_4233	**63.52**	54.79
3.	*HMGR*	1.1.1.88	transcript_43136	transcript_15869	3.99	11.27
4.	*MVK*	2.7.1.36	transcript_30528	transcript_41397	**18.71**	10.96
5.	*PMK*	2.7.4.2	transcript_30845	transcript_20230	**18.22**	2.46
6.	*MVDD*	4.1.1.33	transcript_6471	transcript_39675	**37.89**	19.7
7.	*DXS*	2.2.1.7	transcript_5820	transcript_32850	22.66	49.05
8.	*DXR*	1.1.1.267	transcript_1841	transcript_3020	28.3	160.35
9.	*ISPD*	2.7.7.60	transcript_3471	transcript_14543	**51.25**	12.76
10.	*ISPE*	2.7.1.148	transcript_898	transcript_5328	**109.59**	36.53
11.	*ISPF*	4.6.1.12	transcript_1885	transcript_32702	**450.55**	155.22
12.	*ISPG*	1.17.7.1	transcript_907	transcript_24152	277.77	300.43
13.	*ISPH*	1.17.1.2	transcript_827	transcript_32372	11.68	401.33
14.	*IPPI*	5.3.3.2	transcript_1912	transcript_14601	**196.98**	133.16
15.	*GDPS*	2.5.1.1	transcript_9738	transcript_18602	**14.52**	14.06

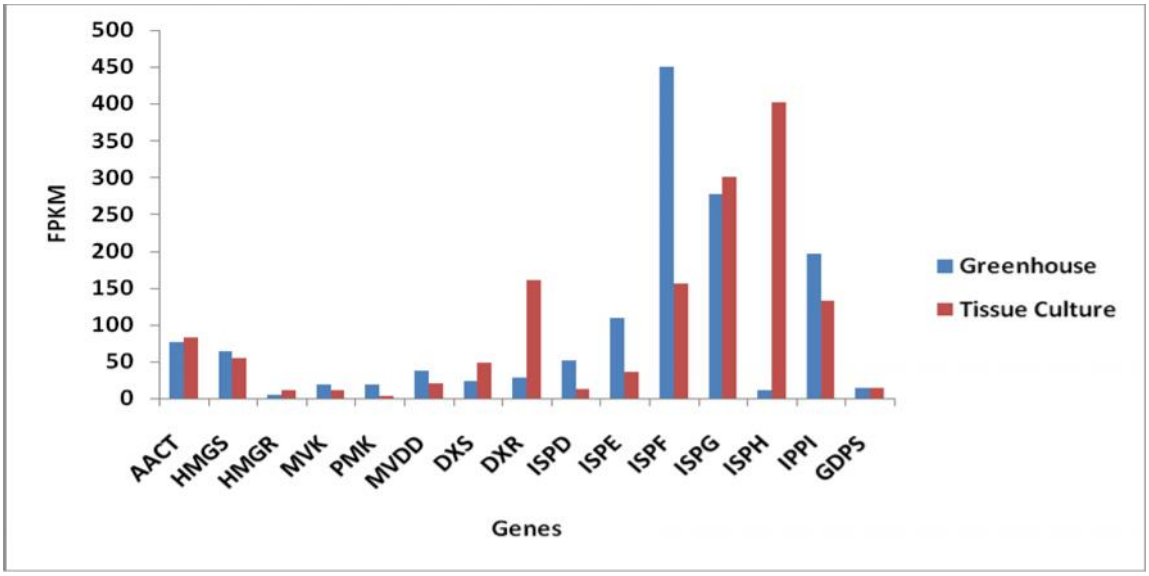


Fig. 10 Transcript abundance of MVA and MEP pathways genes in greenhouse versus tissue transcriptomes

Pathway mapping of transcripts using KEGG

The transcripts were mapped to 337 unique pathways in root transcriptome and 333 unique pathways in shoot transcriptome. 341 number of mapped KOs were found to be involved in the biosynthesis of secondary metabolites for AHSR and 329 for AHSS, respectively. From 57,460 transcripts in SCFG, 8,690 transcripts were assigned KO which were further mapped onto 342 KEGG pathways. Out of 8,690 transcripts, 830 (12.65%) were related to metabolic pathways, 351(5.35%) to the biosynthesis of secondary metabolites. Similarly, in SCTC, out of 43,702 assembled transcripts, which were mapped to 341 KEGG pathways. Out of 6,991 transcripts, 796 (12.86%) were related to metabolic pathways, 341 (5.50%) to the biosynthesis of secondary metabolites, 22 (0.355%) to citrate cycle (TCA cycle). Moreover, the comparative KEGG analysis for within *A. heterophyllum and S. chirayita* datasets will help us to unravel the major enzymes and pathways encoding for important secondary metabolites both species.

Functional classification using NCBI BioSystems

The transcripts with significant matches in the database were assigned to 5 main categories, out of which genetic information processing was the biggest category (1821, 48.07%) in AHSR. On the other hand, metabolism was the biggest category (415, 48.25%) in AHSS. For AHSR, 1,821 transcripts were divided into genetic information processing, including transcription, translation, replication and repair, etc.

Highest number of transcripts belonged to the metabolism category (349, 46.65%) in SCFG transcriptome. On the other hand, for SCTC, genetic information processing formed the biggest category SCTC (280, 47.61%). Network connectivity diagrams were constructed for

isoquinoline alkaloid biosynthesis in root transcriptome of *A. heterophyllum* using in-house developed scripts (Fig. 11, Fig. 12).

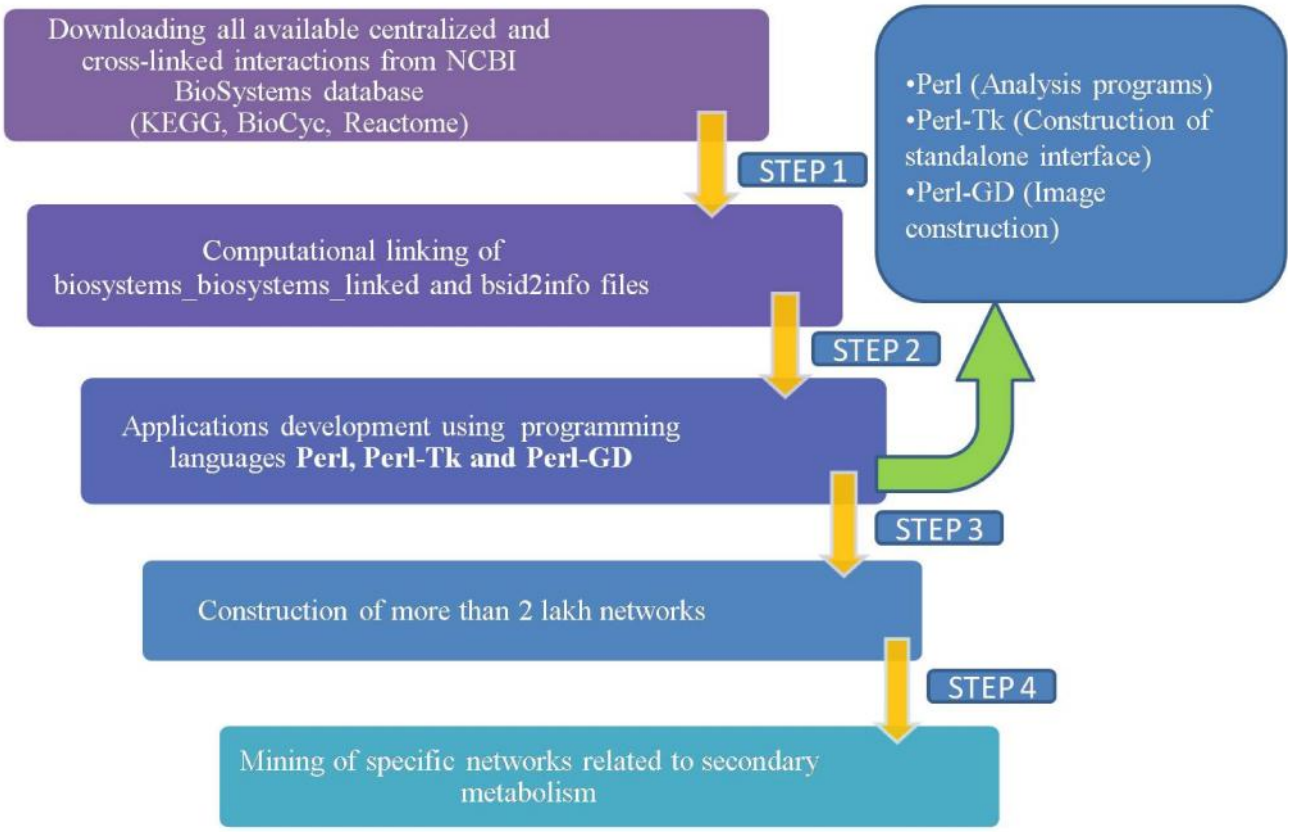

Downloading all available centralized and cross-linked interactions from NCBI BioSystems database (KEGG, BioCyc, Reactome)
STEP 1
Computational linking of biosystems_biosystems_linked and bsid2info files
STEP 2
Applications development using programming languages **Perl, Perl-Tk and Perl-GD**
STEP 3
Construction of more than 2 lakh networks
STEP 4
Mining of specific networks related to secondary metabolism
•Perl (Analysis programs)
•Perl-Tk (Construction of standalone interface)
•Perl-GD (Image construction)

Fig. 11 Flow diagram used for constructing Network connectivity diagrams

Fig. 12 Network connectivity diagram for isoquinoline alkaloids biosynthesis in root transcriptome of *A. heterophyllum*

Conclusion

The GO classification revealed molecular function made up the majority, followed by the biological process and cellular component for all the four transcriptomes i.e AHSR,AHSS,SCFG and SCTC. The COG analysis revealed that 'general function prediction' was the largest group in both tissue samples followed by 'translation, ribosomal structure and biogenesis' in AHSR whereas 'posttranslational modification, protein turnover, chaperones' in AHSS and 'posttranslational modification, protein turnover, chaperones' in AHSR and 'translation, ribosomal structure' in AHSS. For Swertia transcriptomes, COG analysis revealed 'general function prediction' as the largest group followed by 'posttranslational modification, protein turnover, chaperones' and 'translation, ribosomal structure and biogenesis' for both the tissues (SCFG and SCTC). The genes involved in secondary metabolite biosynthesis such as HMGR, MVDD, MVK and HDS showed higher *in silico* transcript abundance in roots as compared to shoots of *A. heterophyllum*. In case of *S. chirayita*, comparative *in silico* analysis of MVA/MEP and phenylpropanoid pathways genes in tissue culture and field grown plant transcriptomes revealed seventeen genes with higher transcript abundance in field grown tissue as compared to tissue culture samples. The network connectivity diagram for isoquinoline alkaloids biosynthesis in root transcriptome of *A. heterophyllum* revealed 77 distinct interacting pathways.

References

[1] J. Ellis, P. Dodds and T. Pryor, "Structure, function and evolution of plant disease resistance genes", *Current Opinion in Plant Biology*, vol. 3, no. 4, pp. 278-284, 2000.

[2] J. L. Dangl and J. D. G. Jones, "Plant pathogens and integrated defence responses to infection," *Nature*, vol. 411, no. 6839, pp. 826–833, 2001.

[3] N. Young, "The genetic architecture of resistance", *Current Opinion in Plant Biology*, vol. 3, no. 4, pp. 285-290, 2000.

[4] W. Sanseverino et al., "PRGdb 2.0: Towards a community-based database model for the analysis of r-genes in plants," *Nucleic Acids Research*, vol. 41, no. D1, pp. D1167–D1171, 2012.

[5] D. Marone, M. Russo, G. Laidò, A. De Leonardis, and A. Mastrangelo, "Plant nucleotide binding Site–Leucine-Rich repeat (NBS-LRR) genes: Active guardians in host defense responses," *International Journal of Molecular Sciences*, vol. 14, no. 4, pp. 7302–7326, 2013.

[6] S. W. Pelletier, R. Aneja, and K. W. Gopinath, "The alkaloids of aconitum heterophyllum wall.: Isolation and characterization," *Phytochemistry*, vol. 7, no. 4, pp. 625–635, 1968.

[7] Z. Wang, J. Wen, J. Xing, and Y. He, "Quantitative determination of diterpenoid alkaloids in four species of Aconitum by HPLC," *Journal of Pharmaceutical and Biomedical Analysis*, vol. 40, no. 4, pp. 1031–1034, 2006.

[8] P. Joshi and V. Dhawan, "*Swertia chirayita* – an overview." *Current Science*, vol. 89, no. 4, pp. 635-640, 2005.

[9] N. Xiao, D. S. Cao, M. F. Zhu, and Q. S. Xu, "Protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences," *Bioinformatics*, vol. 31, no. 11, pp. 1857–1859, 2015.

[10] C. C. Chang and C. J. Lin, "LIBSVM," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 1–27, 2011.

[11] M. L. Bermingham et al., "Application of high-dimensional feature selection: Evaluation for genomic prediction in man," *Scientific Reports*, vol. 5, p. 10312, 2015.

[12] A. Conesa, S. Gotz, J. M. Garcia-Gomez, J. Terol, M. Talon, and M. Robles, "Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research," *Bioinformatics*, vol. 21, no. 18, pp. 3674–3676, 2005.

[13] P. Gahlan et al., "De novo sequencing and characterization of *Picrorhiza kurrooa* transcriptome at two temperatures showed major transcriptome adjustments," *BMC Genomics*, vol. 13, no. 1, p. 126, 2012.

[14] S. Pradhan, N. Bandhiwal, N. Shah, C. Kant, R. Gaur, and S. Bhatia, "Global transcriptome analysis of developing Chickpea (*Cicer arietinum* L.) seeds," *Frontiers in Plant Science*, vol. 5, 2014.

[15] B. Qi, Y. Yang, Y. Yin, M. Xu and H. Li, "De novo sequencing, assembly, and analysis of the Taxodium'Zhongshansa' roots and shoots transcriptome in response to short-term waterlogging", *BMC Plant Biol*, vol. 14, no. 1, 2014.

[16] D. Bhattacharyya, R. Sinha, S. Hazra, R. Datta, and S. Chattopadhyay, "De novo transcriptome analysis using 454 pyrosequencing of the Himalayan Mayapple, *Podophyllum hexandrum*," *BMC Genomics*, vol. 14, no. 1, p. 748, 2013.

[17] R. L. Tatusov, "The COG database: A tool for genome-scale analysis of protein functions and evolution," *Nucleic Acids Research*, vol. 28, no. 1, pp. 33–36, 2000.

[18] B. Li and C. N. Dewey, "RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome," *BMC Bioinformatics*, vol. 12, no. 1, p. 323, 2011.

[19] Y. Moriya, M. Itoh, S. Okuda, A. C. Yoshizawa, and M. Kanehisa, "KAAS: An automatic genome annotation and pathway reconstruction server," *Nucleic Acids Research*, vol. 35, no. Web Server, pp. W182–W185, 2007.

[20] L. Y. Geer et al., "The NCBI BioSystems database," *Nucleic Acids Research*, vol. 38, no. Database, pp. D492–D496, 2009.

[21] C. B. Burge and S. Karlin, "Finding the genes in genomic DNA," *Current Opinion in Structural Biology*, vol. 8, no. 3, pp. 346–354, 1998.

[22] N. Malhotra, V. Kumar, H. Sood, T. R. Singh, and R. S. Chauhan, "Multiple genes of mevalonate and non-mevalonate pathways contribute to high aconites content in an endangered medicinal herb, *Aconitum heterophyllum* wall," *Phytochemistry*, vol. 108, pp. 26–34, 2014.

[23] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of Molecular Biology*, vol. 215, no. 3, pp. 403–410, 1990.

[24] T. Ideker and N. Krogan, "Differential network biology", *Mol Syst Biol*, vol. 8, 2012.

[25] M. A. Gururani, J. Venkatesh, C. P. Upadhyaya, A. Nookaraju, S. K. Pandey, and S. W. Park, "Plant disease resistance genes: Current status and future directions," *Physiological and Molecular Plant Pathology*, vol. 78, pp. 51–65, 2012.

List of Publications

Papers in Peer Reviewed Journals

[1] **Tarun Pal**, Varun Jaiswal, Rajinder S. Chauhan (2016). DRPPP: A machine learning based tool for prediction of disease resistance proteins in plants. *Computers in Biology and Medicine*, 78, 42-48. **(Impact factor: 1.521) Scopus**

[2] **Tarun Pal**, Nikhil Malhotra, Sree Krishna Chanumolu, Rajinder S. Chauhan (2015). Next-generation sequencing (NGS) transcriptomes reveal association of multiple genes and pathways contributing to secondary metabolites accumulation in tuberous roots of *Aconitum heterophyllum* Wall. *Planta*, 242(1), 239-258. **(Impact factor: 3.376) Scopus**

[3] **Tarun Pal**, Jibesh Kumar Padhan, Pawan Kumar, Hemant Sood, Rajinder S. Chauhan (2016). Comparative transcriptomics uncovers differences in photoautotrophic versus photoheterotrophic modes of nutrition in relation to secondary metabolites biosynthesis in *Swertia chirayita*. *Molecular Biology Reports* **(Impact factor: 1.698) (Revision Submitted) Scopus**

[4] Varun Gupta, Nikhil Malhotra, **Tarun Pal**, Rajinder S. Chauhan (2016). Molecular dissection of pathway components unravels atisine biosynthesis in a non-toxic Aconitum species, *A. heterophyllum* Wall. *3Biotech* 6(1), 1-10. **(Impact factor: 0.992) Scopus**

Conference Publications

[1] **Tarun Pal**, Sree Krishna Chanumolu, Varun Jaiswal, Rajinder S. Chauhan (2013). Computational pipeline for analysis of ngs transcriptome data sets in medicinal herbs in *7th Annual Convention of ABAP & International Conference on Plant Biotechnology, Molecular Medicine & Human Health*, University of Delhi, South Campus, October 18-20, 2013. pp 193

[2] **Tarun Pal**, Varun Jaiswal, Sree Krishna Chanumolu, Rajinder S. Chauhan (2014) Computational mining of *Aconitum heterophyllum* transcriptomes for transcription factors controlling atisine biosynthesis in *National Symposium on Advances in Biotechnology for Crop Improvement*, Eternal University, Baru Sahib (H.P.), July 12, 2014. pp 37-38